DOCUMENT RESUME

ED 310 170 TM 013 859

AUTHOR Hunter, John E.
TITLE The Economic Benefits of Personnel Selection Using Ability Tests: A State of the Art Review Including a Detailed Analysis of the Dollar Benefit of U.S. Employment Service Placements and a Critique of the Low-Cutoff Method of Test Use. USES Test Research Report No. 47.
INSTITUTION Employment and Training Administration (DOL), Washington, D.C.; Utah State Dept. of Employment Security, Salt Lake City. Western Test Development Field Center.
PUB DATE 83
NOTE 33p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Ability Identification; *Aptitude Tests; *Cost Effectiveness; Cutting Scores; *Economic Factors; Economic Research; Employment Services; Federal Aid; Job Applicants; Mathematical Formulas; Occupational Tests; *Personnel Selection; State of the Art Reviews; *Test Use
IDENTIFIERS General Aptitude Test Battery; Low Cutoff Method; Placement Tests; United States Employment Service

ABSTRACT

The economic impact of optimal selection using ability tests is far higher than is commonly known. For small organizations, dollar savings from higher productivity can run into millions of dollars a year. This report estimates the potential savings to the Federal Government as an employer as being 15.61 billion dollars per year if tests were given optimal use. If the 4 million placements per year made by the United States Employment Service made optimal use of the General Aptitude Test Battery, the potential increase in work force productivity among the employers who hire through the service would come to 79.36 billion dollars per year. However, this would probably require an increase in Employment Service funding of about 8.75 million dollars per year. Departures from optimal use of tests can be shown to eliminate as much as 84% of these savings. The principal problem is the use of the low-cutoff method of hiring randomly from all who pass some minimal test level. Optimal use of tests can be shown to provide benefits other than reduced labor costs, including a reduction in special administrative problems, an increase in the number of workers with promotion potential, and increases in the quality as well as the quantity of work. Five tables provide supporting figures. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED310170

USES TEST RESEARCH REPORT NO. 47

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

THE ECONOMIC BENEFITS OF PERSONNEL SELECTION USING ABILITY TESTS: A STATE OF THE ART REVIEW INCLUDING A DETAILED ANALYSIS OF THE DOLLAR BENEFIT OF U.S. EMPLOYMENT SERVICE PLACEMENTS AND A CRITIQUE OF THE LOW-CUTOFF METHOD OF TEST USE

DIVISION OF COUNSELING AND TEST DEVELOPMENT
EMPLOYMENT AND TRAINING ADMINISTRATION
U.S. DEPARTMENT OF LABOR
WASHINGTON, D.C. 20213

1983

TM013859

BEST COPY AVAILABLE

TABLE OF CONTENTS

	Page
LIST OF TABLES	i
ACKNOWLEDGMENT	iii
ABSTRACT	v
INTRODUCTION	1
Overview of the Report	2
BENEFITS OF OPTIMAL USE OF TESTS	4
The Basic Utility Formula	5
Dollar Savings in Quantity of Work	6
Wages, Output, and S_y	6
An Example: The Federal Government	8
Reduction in Administrative Problems	9
Increasing the Promotion Talent Pool	10
Noncompensatory Utility: The Effect of Quality of Work	12
Benefits of Test Use to Applicants	12
HIGH PRODUCTIVITY VERSUS RACIAL BALANCE	13
Adverse Impact and Test Fairness	13
Racial Differences on Different Abilities	14
Savings Losses for Nonoptimal "Models of Fair Test Use"	14
Economic Disaster: The Low-Cutoff Procedure	17
POTENTIAL AND ACTUAL ECONOMIC BENEFITS FOR ORGANIZATIONS WHICH HIRE THROUGH THE U.S. EMPLOYMENT SERVICE	20
Potential Economic Benefit to Employers	21
Actual Productivity Gains for Placements	21
CONCLUSION	23
Reference Notes	24
References	25

LIST OF TABLES

Table Number	Page	
1	10	The percentage of very poor workers selected with optimal use of ability tests as a function of the selection ratio (in percentage form) and the validity coefficient.
2	11	The percentage of workers with promotion potential selected given optimal use of ability tests as a function of validity and selection ratio.
3	16	Summary of the implications of different models of fair use of tests in terms of economic productivity and in terms of minority hiring; adapted from Hunter, Schmidt, and Rauschenberger (1977).
4	18	A comparison of the productivity losses and relative minority hiring rates for 5 methods of using ability tests for selection (selection ratio is 10 percent, validity is .50). The four methods abbreviated are those considered in Table 3: C=Cleary, T=Thorndike, D-C=Darlington-Cole, and Q=quota. The minority baseline is assumed to be 20 percent.
5	20	A comparison of four different methods of using an ability test to select entry level workers into the Federal Government (validity of .50, selection ratio of 10 percent, minority baseline of 20 percent).

ACKNOWLEDGMENT

The United States Employment Service conducts a test research program for developing testing tools useful in vocational counseling and placement.

The purpose of this series of reports is to provide results of significant test research projects as they are completed. These reports will be of interest to users of USES tests and to test research personnel in state agencies and other organizations.

Special thanks go to John Hawk of the U.S. Employment Service and Ron Boese of the North Carolina Employment Security Commission for assembling most of the government statistics used in this paper. Mr. Boese not only put together the statistics on wages, but also analyzed them by job family so that accurate wage figures could be obtained for the U.S. Employment Service whose job distribution differs from that of the nation as a whole.

This report was written by Dr. John E. Hunter, Michigan State University, under contract to the Utah Department of Employment Security, Western Test Development Field Center, Salt Lake City, Utah. The report was prepared for printing by staff of the Western Test Development Field Center.

ABSTRACT

The economic impact of optimal selection using ability tests is far higher than is commonly known. For small organizations, dollar savings from higher productivity can run into millions of dollars a year. This report estimates the potential savings to the Federal Government as an employer as being $15.61 billion per year if tests were given optimal use. If the 4 million placements per year made by the U.S. Employment Service made optimal use of the General Aptitude Test Battery, the potential increase in work force productivity among the employers who hire through the service would come to $79.36 billion per year. Departures from optimal use of tests can be shown to eliminate as much as 84 percent of these savings. The principal problem is the use of the low-cutoff method of hiring randomly from all who pass some minimal test level. Optimal use of tests can be shown to provide benefits other than reduced labor costs including a reduction in special administrative problems, an increase in the number of workers with promotion potential, and increases in the quality as well as quantity of work.

INTRODUCTION

Periods of high inflation sharply etch the need for high productivity. If a government agency is to maintain its level of service during a year of 20 percent inflation with a budget only 10 percent higher, then it must increase productivity by 10 percent. Yet the issue of productivity is just as much present in noninflationary times. If an agency is to maintain its level of service during stationary times, it must do nothing to reduce its level of productivity.

The issue is similar for private industry. The present automobile crisis has clearly revealed the fact that modern corporations are not only competing with foreign corporations for foreign markets but for our domestic markets as well. In the past we have maintained our competitive edge by matching low labor prices in foreign corporations with high levels of productivity. Thus any act which reduces productivity can have a disastrous impact on both foreign and domestic sales.

One crucial element in the maintenance of high productivity is to select people with high ability for their jobs. For most jobs, the only presently known predictive devices with high validity are cognitive ability tests. Recent work on validity generalization (Schmidt and Hunter, 1977; Schmidt, Hunter, Pearlman, and Shane, 1979; Pearlman, Schmidt, and Hunter, 1980; Schmidt, Gast-Rosenberg, and Hunter, 1980; Schmidt, Hunter, and Kaplan, in press; Hunter [Note 1]) has shown that most findings of low validity are due to artifacts of modern empirical studies, mainly statistical error due to small sample size. Similar and broader conclusions follow from reanalyses of past meta-analysis by Lent, Aurbach, and Levin, (1971) as noted by Hunter (1980) and from inspection of unpublished, large-sample studies done in the U.S. Army by Helme, Gibson, and Brogden (Note 2); in Hunter (1980) and in Schmidt, Hunter, and Pearlman (in press).

High test validity translates into considerable dollar savings for most organizations. Hunter (Note 3) recently estimated that if the Philadelphia police department were to drop their use of a cognitive ability test to select entry level officers, it would cost the city over $170 million over a ten-year period. Schmidt, Hunter, McKenzie, and Muldrow (1979) provide figures which show that over a ten-year period, the Federal Government would save $376 million if computer programmers were selected using the Programmer Aptitude Test (PAT). The corresponding figure for the economy as a whole would be $6.22 billion.

The impact of cognitive tests on productivity can be estimated at a national level. Hunter and Schmidt (in press) formed a utility model of the national economy. Gains are not as spectacular at the national level as would be predicted from findings for single organizations because high ability people

who are not selected for crucial top level jobs will be available for lower level jobs and will bring higher productivity to such jobs. However, even with this cancellation, Hunter and Schmidt estimate that productivity differences between complete use or complete disuse of cognitive ability tests would amount to $80 billion per year. That is, productivity differences due to use or nonuse of present tests would be about as great as total corporate profits or about 20 percent of the total Federal budget.

To replace the use of cognitive ability tests by any instrument of lower validity would be to incur very great economic costs. Moreover, these costs fall on everyone, whatever their sex or group affiliation.

Overview of the report

This paper will be written in three parts: (1) the economic benefits of optimal use of tests in personnel selection, (2) the reduction in benefits from various strategies for obtaining racial balance, and (3) the current and potential benefits to business from U.S. Employment Service placements.

Optimum use of tests for personnel selection depends on two things: use of the most predictive ability tests to select for a given job, and selection of the top scorers on the test. Given optimum selection, the benefits of using tests are very great. For the typical employer using the U.S. Employment Service, the benefits of using tests to select given optimal usage would be about 33 percent of labor costs for the jobs in question. The potential savings for one year's hires for the Federal Government would be about $15.61 billion. Furthermore, this is just the savings that derives from increased productivity due to high average productivity. There are other benefits as well. Very poor workers create special administrative problems. The very best workers constitute the ideal pool for future promotions.

Optimal selection using current tests can be shown to cut the number of very poor workers drastically; from 10 percent to 1 percent of the work force for the typical U.S. Employment Service employer. Optimal use of tests greatly increases the top-talent pool; from 10 percent to 31 percent for the typical U.S. Employment Service employer. Finally there are differences in quality of work as well as quantity of work. These are more difficult to evaluate in dollars, but they are often very important. For example, poor quality work may cost the employer a customer. Poor policemen who fail to catch a criminal not only generate an increased number of crime reports to be investigated and reported (i.e. an increase in the quantity of work required), but also mean an increase in the total amount of crime.

Optimum use of tests even has some benefit for the applicant. Low-ability applicants are likely to do poorly on the job. If they are hired, they have little likelihood of being above average in performance and hence little likelihood of having a positive self-concept in regards to work. On the

other hand, low-ability applicants run a much higher risk of being in the bottom 10 percent of the work force and hence living under constant "harassment" from supervisors and co-workers.

Optimum use of tests requires that applicants be hired from the top down on ability test score. This is known to lead to a somewhat lower proportion of hiring for nonoptimum minority group members. Thus many have recommended use of tests in order to increase the number of minority persons hired. This always leads to an increase in the number of lower ability persons hired and hence to a lowering of economic savings. However, the loss of utility is higher for some methods of generating racial balance than others. Optimum use of tests within ethnic or racial groups leads to the lowest loss in utility. Hunter, Schmidt, and Rauschenberger (1977) showed that even population quotas rarely lead to a loss of more than 10 percent of the savings generated by use of tests for selection. However, the use of low-cutoff scores (with random hiring above those cutoffs) eliminates most of the benefit of testing; depending on just how low the bottom cutoff is set. If applicants are hired at random from the top two-thirds of the ability distribution, then the typical employer using the U.S. Employment Service loses 70 percent of the savings that accrue from hiring on the basis of ability. If hiring is at random from the top 80 percent of the distribution, then 84 percent of the savings are lost. Thus if the Federal Government were to replace optimal test use by the use of low cutoffs, the savings due to increased productivity would drop from $15.61 billion per year to only $2.50 billion per year. That is, nonoptimal use of tests for selection would cost the government $13.11 billion per year.

Ironically, the disastrous effects of the use of low cutoffs for selection, instead of ranking, are not matched by corresponding gains in minority employment. The use of low cutoffs leads to hiring fewer minority applicants than would be true of quotas. Thus the low-cutoff procedure is a disaster for both employers and for minority applicants.

The U.S. Employment Service placed 4,022,019 applicants in jobs in 1980. If the U.S. Employment Service had made optimal use of their test battery, the GATB, then the potential savings to employers would have been $79.36 billion in 1980. However, the U.S. Employment Service does not use tests in an optimal way. There are three drawbacks to present test use:

(1) Tests are used in fewer than 10 percent of placements.
(2) Prediction equations are based on small sample studies instead of validity generalization.
(3) Recommendations are made on the basis of low cutoffs instead of ranking.

Thus it is likely that the utility gains to employers using the U.S. Employment Service are not $79.36 billion, but only $1.73 billion. That is, shortcomings in the present U.S. Employment Service procedures are costing

American business $77.63 billion per year; i.e. employers who hire through the U.S. Employment Service are losing 98 percent of what they might have gained in increased productivity.

To go from the present wasteful disregard of optimal use of the GATB to optimal use would require two different kinds of administrative change. To give the test to more applicants would require an increase in personnel at the local office. However, the GATB can be given in groups by low-level clerical personnel, and hence this increase in cost is far less than the corresponding gains to the business community. The second kind of change is free; it requires only a change in how test scores are used for applicants who would take the test anyway. The Washington office would merely need to issue new guidelines for test use:

(1) Use the prediction equations based on validity generalization.
(2) Hire by ranking instead of recommending any who exceed the medium (M) or high (H) cutoffs (i.e. hiring at random from among the top 80 percent).

Changing how test scores are used would save employers $7.94 billion per year. This may be far less than the potential of $79.36 billion per year, but is also far higher than the current $1.73 billion per year. That is, switching from low cutoffs to ranking for recommendations would save U.S. Employment Service employers $6.21 billion per year.

BENEFITS OF OPTIMAL USE OF TESTS

The classic formulas for the dollar benefit of using ability tests to select applicants for employment were derived first by Brogden (1946, 1949) and by Cronbach and Gleser (1965) in very complex form. Much simpler and more straightforward derivations of these formulas were presented in Hunter and Schmidt (in press) and in Schmidt, Hunter, McKenzie, and Muldrow (1979). They show that the basic equation is the regression equation for production in dollar terms onto test score. This equation assumes only linearity of that regression. This linearity assumption has been verified by the examination of thousands of empirical studies for nonlinearity. These cumulative studies are reviewed in Hunter and Schmidt (in press) and by Schmidt et al. (1979). In particular, Hawk (1970) looked at 3,303 relationships between GATB aptitude scores and job proficiency for nonlinear relationships. He found statistical evidence for nonlinearity at exactly the chance level. Thus all the cumulative studies are agreed: There are no nonlinear relationships between test scores and job proficiency in the present job market.

The basic utility formula

The average gain from the use of an ability test to applicants for the job is the difference between average performance for those selected using the test and average performance for those selected using whatever alternative procedure the employer would use (procedures known from empirical studies to be little better than random hiring). Average gain is also called average marginal utility or average utility for short and is denoted $\bar{U}$. The formula for average gain is

$$\bar{U} = r_{xy} \; s_y \; \bar{x} \qquad \text{Equation (1)}$$

where:

r_{xy} = the validity of the test for predicting true job performance in the applicant pool

s_y = the standard deviation of true job performance in dollars in the applicant pool

$\bar{x}$ = the average test score of those selected in applicant pool standard scores

The validity coefficient r_{xy} is the correlation between test score and true job performance calculated across all applicants. If this validity coefficient is to be estimated by an observed correlation from a typical validation study, then the observed correlation must be corrected for two sources of systematic error in validation studies: error of measurement of job performance and reduction in correlation due to restriction in range. These correction formulas have been known for many years (see for example, Schmidt, Hunter, and Urry, 1976), and are also embedded in validity generalization approaches to test validation (Schmidt and Hunter, 1977; Schmidt, Hunter, Pearlman, and Shane, 1979; Pearlman, Schmidt, and Hunter, 1980; Schmidt, Gast-Rosenberg, and Hunter, 1980; Schmidt, Hunter, and Pearlman, in press; Hunter [Note 4]; Callender and Osburn, 1980).

The standard deviation (s_y) is the standard deviation of job performance in dollars. This is the number that is most difficult to obtain in practice. It is so difficult that even over a 35-year span, only a handful of utility studies were seen in the empirical literature. However, cumulative work and theoretical progress have provided an alternative to cost accounting to estimate this number. This alternative strategy will be presented in the section of the paper on wages and output.

The mean test score ($\bar{x}$) varies according to the number of applicants selected. The smaller the percentage of applicants to be selected, the higher the average test score of those chosen. Since test-score distributions are nearly always normally distributed in the applicant

population, the number represented by $\bar{x}$ can be calculated from the selection ratio (SR). First, use a normal curve table to determine the cutoff score required to select the top SR percentage of the population. Denote that cutoff score by c and express SR as a decimal proportion. Then the mean test score in standard score form is given by

$$\bar{x} = \frac{\phi(c)}{SR} \qquad \text{Equation (2)}$$

where ϕ is the normal density function or normal curve "ordinate."

Dollar savings in quantity of work

If performance in dollar terms is given in annual levels, then the average utility formula gives dollars saved per year per person. Thus the total savings in a year's hires must be aggregated along two dimensions: job tenure and number of hires. For example, if a poor worker is hired, then the employer must suffer the loss over the entire time that the worker is with the organization. Therefore, the average savings for a hire is the average savings for each year multiplied by the number of years that the worker stays. Note that tenure is defined as the number of years with the organization, not the number of years at the job for which the applicant was hired. If the worker is promoted, then his productivity is even higher. That is, to multiply savings by job tenure is to underestimate the value of those workers who are subsequently promoted. Let T be the average job tenure, i.e. the average number of years that the selected applicant stays with the organization. Let N be the number of persons hired in a given year. Then the total utility of a year's hires, denoted U, is given by the product of N times T times average gain $\bar{U}$, i.e. by

$$U = N T \bar{U} = N T r_{xy} s_y \bar{x} \qquad \text{Equation (3)}$$

Wages, output, and s_y

It was once thought that the standard deviation s_y could only be estimated by cost accounting. However, attempts to do so proved very frustrating. Not only is cost accounting very expensive, but it involves many arbitrary decisions and hence has a considerable degree of error. Thus most of the empirical studies located by Hunter and Schmidt (in press; see also Schmidt, Hunter, Muldrow, and McKenzie, 1979) considered only partial measures of dollar value such as savings in training costs, or dollars saved by reduction in accidents, or other administratively convenient values. In order to cumulate these results across type of job, across years, and even across national monetary units, they expressed the empirical standard deviations in ratio to the average wage of the worker studied.

To get around the problem of partial utility, they also invented a new method of estimating s_y. For certain jobs, there is a person in the organization who knows what it would cost to replace a given worker by an outside firm or consultant. These experts can then be queried as to what it would cost to replace an average worker, a worker who is better than 85 percent of his co-workers, or a worker who is worse than 85 percent of his co-workers. These numbers can be compared for consistency and combined to provide an estimate of s_y. The estimates for different judges can then be compared to see if there is a consistent market value. If the estimates are generally similar, then the mean judgment can be used as the final estimate, and the variance across judges can be used to estimate the error in the average judgment.

The values compiled and presented in Hunter and Schmidt (in press, or in Schmidt et al., 1979) centered about the value of 40 percent of wages. Considerable work done since then but not yet published, tends to verify that value (see Mack, Schmidt, and Hunter [Note 5], for one such study; a review paper is now being written). Thus, we are now convinced that the number (.40 annual wage) can be taken for the baseline estimate of s_y. The jobs which vary from this value are in the direction of being much higher. For example, the difference in dollars recovered by income tax investigators can run into the hundreds of thousands of dollars and hence be far greater than the wage paid. Also people who supervise certain critical machines or operations in a steel plant can make errors that cost hundreds of thousands of dollars. Thus a person who makes very many errors can cost far more than his total wages for a lifetime.

For some time we were puzzled by the fact that our standard deviation is 40 percent of wages. If you go 2.5 standard deviations below the mean on a variable whose standard deviation is 40 percent of the mean, you get 0. Could it be that work performance typically goes all the way down to 0? However, this is erroneous thinking. The standard deviation s_y is not the standard deviation of wages; in most jobs all workers are paid the same and the standard deviation in wages is 0. Rather s_y is the standard deviation in work output, i.e. the differences in the value of the work produced.

What is the relationship between wage and output? We knew that for most businesses serving households (such as plumbers or TV repair), the employer usually charges the customer about twice what the worker is paid. Thus wages tend to be about half the worth of the product; the other half representing overhead, materials, the labor of others, etc. This figure was confirmed in national economic statistics. Thus to say that s_y is about 40 percent of wages is to say that s_y is about 20 percent of output.

With this realization, we noted that our basic finding can be phrased in a much more prosaic way: Workers two standard deviations above the mean produce about 40 percent more than average. Workers two standard deviations

below the mean produce about 40 percent less than average. Thus the ratio in output of top to bottom workers is about

$$\frac{\text{Top Output}}{\text{Bottom Output}} = \frac{1.40}{.60} = 2.33$$

Thus our findings can be stated as follows: For the typical job, the top workers produce about twice as much as the bottom workers. Managers we have questioned find this figure of 2 to 1 to be quite plausible.

In using the estimate of s_y equal 40 percent of annual wage, the key question is this. Is the ratio of productivity between top and bottom workers on this job more than two to one, less than two to one, or about two to one? The answer to this question shows the direction of error in using the baseline figure for the average job as an estimate for any given job.

An example: the Federal Government

What is the potential annual savings in labor cost if the Federal Government were to use tests in an optimal way to select new workers? Employment and Earnings (DOL, 1980) shows that the government employs three million workers with an average job tenure of 6.52 years. Thus there are about 460,000 new workers hired each year. The average annual wage is $13,598. Informal inquiries at the Office of Personnel Management suggested that there are usually at least 10 applicants for each government job opening. Thus the selection ratio is about 10 percent and the test score cutoff should be about $c = 1.28$ standard deviations above the mean, which yields $x = 1.76$.

The test-validity estimation is more complicated since it uses a validity generalization recently completed on the 415 validation studies compiled by the U.S. Employment Service. According to Hunter (Note 5), most government jobs fall into JOBFAM categories 2 and 3. Thus, the GATB cognitive-aptitude composite score would have validity .55 in selecting government workers.

The total savings in labor costs represented by optimal use of tests for one year is then given by

$$U = N\,T\,r_{xy}\,s_y\,\bar{x}$$

$$= (460{,}000)(6.52)(.55)\quad(.40)(13{,}598)\quad(1.76)$$

$$= \$15.61 \text{ billion.}$$

A figure of $16 billion may sound large, but it must be figured against the total amount of money being paid in wages. The total amount of money being spent on one-year's hires is the number of persons times the number of years times the average wage, i.e.

$$\text{Total Spent} = N\ T\ \text{Wage} = (460{,}000)(6.52)(13{,}598)$$

$$= \$40.78 \text{ billion.}$$

Thus the ratio of savings to spending is given by

$$\frac{\text{Savings}}{\text{Expenditures}} = \frac{15.61}{40.78} = 38 \text{ percent}$$

The large values which consistently arise in utility computations seem surprising because most people in personnel work do not think in terms of aggregate labor costs. Any process which can save as much as 30 percent of labor costs will save millions of dollars even in small organizations.

Reduction in administrative problems

Very poor workers not only produce less, they also create special administrative problems. They require more monitoring and they frequently become angry or resentful over what they perceive to be constant "harassment." They make mistakes which require make-up work. They may upset customers or co-workers. They tend to be safety risks. For convenience, let us assume that the "very poor" workers are those who would fall in the bottom 10 percent on job performance under random selection. Then reduction in the number of very poor workers would have utility above and beyond that which is measured by the utility equations for quantity of work.

Optimal use of tests can drastically reduce the number of very poor workers (Taylor and Russell, 1939). The extent of reduction depends on the validity of the test and on the selection ratio. The higher the validity coefficient and the more extreme the selection ratio, the greater the reduction in the number of very poor workers. Table 1 presents the reduction for a sample of validity values and a sample of selection ratios.

Table 1. The percentage of very poor workers selected with optimal use of ability tests as a function of the selection ratio (in percentage form) and the validity coefficient.

		Selection Ratio				
		80	50	20	10	5
Validity	.30	8.0	5.8	3.9	3.0	2.4
	.40	7.2	4.6	2.4	1.7	1.2
	.50	6.3	3.4	1.4	.7	.4
	.60	5.3	2.3	.7	.2	.1
	.70	4.4	1.3	.2	.0	.0

An employer who uses the U.S. Employment Service test could have a validity of .50 and a selection ratio of 10 percent. Thus according to Table 1, if the U.S. Employment Service made optimal use of the GATB, the employer using this service could reduce the number of very poor workers from 10 percent to .7 percent. That is, optimal test use would permit reduction by a factor of 14.3, and elimination of over 90 percent of the special administrative problems associated with such workers.

The Federal Government test has a validity of .55 (which is not given in Table 1) and a selection ratio of 10 percent. The corresponding percentage of very poor workers after selection by ability test is .4 percent. Thus the number of very poor workers under optimal test use is reduced by a factor 25. The number of special problems is reduced to only 4 percent of what it would have been.

Increasing the promotion talent pool

Most organizations rely on promotion from within to fill higher level jobs. Thus the quality of personnel at such higher jobs at one point in time is a function of the number of highly talented workers at lower level jobs at an earlier point in time. Thus, over time, the quality of entry level hires

spreads upwards through the organization. It is crucial then to make sure that the pool of entry level workers contains a subset of top talent for promotion. Taylor and Russell (1939) showed that optimal use of a valid ability test can greatly increase the percentage of workers selected who lie in the top-talent category.

To quantify the impact of an ability test, we must define the phrase "top talent." Let us define top talent to mean workers in the top 10 percent of the performance dimension under random selection. Thus under random selection, there would be 10 percent of the work force in the promotion level pool. Table 2 presents the percent of top talent selected under optimal use of ability tests for a sample of validity values and a sample of selection ratios.

Table 2. The percentage of workers with promotion potential selected given optimal use of ability tests as a function of validity and selection ratio.

		Selection Ratio				
		80	50	20	10	5
Validity	.30	11.5	14.3	18.7	21.8	24.5
	.40	12.1	15.7	22.1	26.8	31.6
	.50	12.3	16.9	26.1	32.6	39.4
	.60	12.3	18.1	30.2	39.4	49.4
	.70	12.3	19.2	35.2	47.2	59.5

For an employer who uses the U.S. Employment Service, the validity coefficient is .50 and the selection ratio is about 10 percent. Thus if the U.S. Employment Service made optimal use of the GATB, the increase in top talent would be from 10 percent of the work force to 32.6 percent. That is, optimal use of the GATB would more than triple the number of workers suitable for promotion.

For the Federal Government, the validity is .55 and the selection ratio is 10 percent. Thus the percentage of top talent given optimal use of ability

tests would increase the number of workers with promotion potential from 10 percent to 36.7 percent; i.e. almost quadruple the top-talent pool.

Noncompensatory utility: the effect of quality of work

Hunter and Schmidt (Note 6) noted that the conventional utility formulas of equations (1) and (2) refer only to quantity of work. It is always possible to compensate for differences in quantity of work by using more workers. However, they note that there are many instances in which quality of work is critical. In such instances it is not possible to compensate for lower quality of work by hiring more workers.

It is difficult to quantify the impact of differences in quality since the effects differ from situation to situation. However, it is important that such effects be considered in any particular employment situation, since the presence of such effects may override considerations of quantity and hence may rule out any alternative to optimal test use.

One of the examples presented by Hunter and Schmidt concerns police work. Consider the detail responsible for rape control. Suppose that a mediocre detective is only half as likely to make use of clues from a rape report as a top detective. Then a rapist will go twice as long before being caught. This means that twice as many rape reports will be required of the department. The department could compensate for using poorer detectives by hiring twice as many of them for rape control; this is the meaning of the usual utility formula. However, even using twice as many detectives will not compensate for the difference in quality of work: The community will still suffer twice as many rapes.

Benefits of test use to applicants

Most people consider selection solely in terms of whether the applicant gets the job or not. Few consider the implications of being hired for the applicant. It is true that being hired means having a job, but work means far more than this to most workers. Sociologists have long noted that self-concept is frequently tied to work performance. In particular, people tend to feel self-confident if they do well at their work. Thus feelings of self- confidence will depend on the extent to which the worker can surpass the standard for good work at the job in which he is placed.

Standards differ from job to job. For illustrative purposes, assume that the standard is average performance for workers randomly selected to the job. Then under random selection, half the workers will have a positive self-concept with respect to work. The Taylor and Russell (1939) procedures show that optimal selection using ability tests will greatly increase the number of workers who will feel good about their work. For the typical employer

using the U.S. Employment Service, the test validity is .50 and the selection ratio is 10 percent, and the proportion of selected workers who will feel good about their work is increased from 50 percent to 84 percent.

The problem is even more critical for poor workers. A very poor worker is constantly in trouble with his supervisor and is likely to be angry much of the time and very unhappy at work. Under random selection, the proportion of workers who suffer such harassment is 10 percent. If the U.S. Employment Service were to use optimal selection with the GATB, then the frequency of such high stress placements would decrease from 10 percent to .7 percent, i.e. decrease by a factor of over 10.

Workers are much more likely to be happy if they are placed in jobs where they do well. Optimal use of ability tests for placement greatly increase the probability of such placement.

HIGH PRODUCTIVITY VERSUS RACIAL BALANCE

Adverse impact and test fairness

In the case of cognitive tests, the problem is large differences in the mean ability scores of different racial groups. There is about a one standard-deviation difference not only on verbal ability, but on numerical ability and spatial ability as well. Since black applicants score lower on cognitive ability tests, they are more likely to fall below selection-cutoff scores than are white applicants. For example, if a test is used to select at a level equivalent to the top half among white applicants, it will select only the top 16 percent of the black applicants. This difference is what the courts call "adverse impact."

Fifteen years ago, the elimination of adverse impact seemed a straightforward though arduous task, just adjust the tests. Assuming that there are no differences between racial groups in developed ability, the differences showing on the test would mean that the tests are unfair to black applicants. If the content that is culturally biased could be removed from the test, then not only would adverse impact vanish but the validity of the test would increase. Moreover, a test which is culturally unfair to blacks would probably be culturally unfair to disadvantaged whites as well.

However, the empirical evidence of the last fifteen years has been unkind to this hypothesis. Evidence showing that single-group validity is an artifact of small sample sizes (Schmidt, Berner, and Hunter, 1973; O'Connor, Wexley, and Alexander, 1975; Boehm, 1977; Katzell and Dyer, 1977) has shown that any test valid for one racial group is valid for the other. Evidence showing differential validity to be an artifact of small sample size (Bartlett,

Bobko, Hannan, and Mosier, 1978; Hunter, Schmidt, and Hunter, 1979) has shown that validity is actually equal for the two groups. Finally, there is the cumulation of evidence testing directly for cultural bias, results of which are consistently in the opposite direction to that predicted by the test-bias hypothesis. If test scores for blacks were lower than their true ability scores, then their job performance would be higher than their test scores would predict. But in fact regression lines for black applicants are either below or equal to the regression lines for white applicants (review studies cited in Schmidt and Hunter, 1980).

The evidence is clear: The difference in ability test scores is mirrored by a corresponding difference in academic achievement and in performance on the job. Thus the difference in mean test scores reflects a real difference in mean ability. If the difference is the causal result of poverty and hardship, then it will vanish over time. However, since the difference represents a real difference in ability at the time when tests are taken, there will be no reduction in adverse impact produced by the construction of better tests. In fact, better tests are somewhat more reliable and hence show slightly larger adverse impact.

Racial differences on different abilities

Racial differences are not the same on all abilities. The GATB can be scored in terms of three-abilities composites: cognitive ability, perceptual ability, and psychomotor ability. The differences between the means for blacks and whites are .84, .86, and .29 standard deviations for cognitive, perceptual, and psychomotor ability respectively. That is, there is a much larger difference on cognitive ability than on psychomotor ability. This is very important, because Hunter (Note 5) has shown that for many jobs psychomotor ability is a much better predictor than is cognitive ability. If psychomotor ability is used to select for such jobs, then there will be much less adverse impact than is familiar to the testing literature. For example, if the cutoff score is set to select the top 50 percent of white applicants, then for psychomotor ability the percentage of blacks who would be selected is 39 percent (as opposed to 16 percent for cognitive ability). Thus there is much less reduction in labor-cost savings if alternative "models of test fairness" are used to set quotas. On the other hand, the reduction in savings from the use of random hiring above low cutoffs is just as disastrous for psychomotor as for cognitive ability, and even more painful since it is even less justified.

Savings losses for nonoptimal "models of fair test use"

Once it became clear that test scores are fair to blacks as individuals, the argument within the technical literature shifted to fair use of tests rather than test fairness. This difference in terminology represents a shift from

the scientific issue to fairness of test scores to the ethical issue of racial balance. Hunter and Schmidt (1976) identified four such "models of fair use of tests": the Cleary model (Cleary and Hilton, 1968); the Thorndike (1971) model; the Darlington-Cole model (Darlington, 1971; Definition 3; Cole, 1973); and the quota model. Hunter and Schmidt showed that all four definitions revolved around ethical issues rather than scientific issues; i.e. they are concerned with racial balance rather than with fairness of test scores as measures of ability.

They also showed that all four models could be viewed in terms of quotas for blacks. The Cleary model asserts that the proper quota for blacks is that based on ability to do the job. If 10 percent of the applicants are to be hired, then the quota for blacks would be the percentage of blacks who are in the top 10 percent on ability. The Thorndike model asserts that the proper quota for blacks is the percentage of blacks that would have been selected had the test had perfect validity. The Darlington-Cole model also links the proper quota to the percentage hired. If 10 percent of the applicants are to be hired, then they define "success" on the jobs as being in the top 10 percent in job performance. They then set the quota of blacks so that the conditional probability of being hired if actually successful is the same for blacks as for whites. The quota model asserts that the proper quota for blacks is the percentage of blacks in the population. The four models are listed here in the order of the size of the quota that they define for blacks, with the Cleary model setting the lowest quota and the quota model setting the highest.

Hunter, Schmidt, and Rauschenberger (1977) showed in their appendix that a test can always be scored to make it "fair" according to any of the four models. They derived the number of points that would have to be added to black test scores to make the test fair by each definition. Thus one need not write a new test to shift from one definition to another (which is a good thing since all content-valid tests have proved to be fair only according to the Cleary definition). Thus the four models can be viewed as alternative ways to score tests rather than alternative procedures for assessing tests.

These four methods can be assessed on scientific grounds. Which method produces the more valid scoring? The empirical evidence here is clear. The Cleary method of scoring maximizes the validity of the test. Adding points to achieve racial balance reduces the scientific worth of the instrument.

These four methods can also be evaluated economically. Which method produces the work force with highest productivity? Again the empirical evidence is clear. The Cleary method maximizes the mean productivity of the group of applicants hired. However, one can ask about an economic tradeoff: How much money should an organization be willing to lose in order to achieve racial balance? That, of course, is a matter of values. On the other hand, it is a matter of science to calculate the cost of using each of these scoring

methods to select applicants. This was done by Hunter, Schmidt, and Rauschenberger (1977) and their results were used to construct Table 3.

Table 3. Summary of the implications of different models of fair use of tests in terms of economic productivity and in terms of minority hiring; adapted from Hunter, Schmidt, and Rauschenberger (1977).

Table 3a. Results for a selection ratio of 10 percent (with validity of .50 and a minority baseline of 20 percent).

	Models or scoring methods			
	Cleary	Thorndike	Darlington-Cole	Quota
Percent Savings Lost	0	1	3	5
Percent Minority Hired	1.5	4.4	6.8	10.0

Table 3b. Results for a selection ratio of 50 percent (with validity of .50 and a minority baseline of 20 percent).

	Models or scoring methods			
	Cleary	Thorndike	Darlington-Cole	Quota
Percent Savings Lost	0	2	4	7
Percent Minority Hired	21	34	41	50

Table 3 presents a summary of the findings of Hunter, Schmidt, and Rauschenberger (1977) for a validity of .50 (the general finding in all major job categories according to Hunter, Note 5) and a minority baseline of 20 percent representing 10 percent black and 10 percent Hispanic. Table 3a

shows the results for a selection ratio of 10 percent, and Table 3b shows the results for a selection ratio of 50 percent. Each table shows the same stark tradeoff; that method (quota method) which maximizes minority hiring also maximizes the extent of economic loss to the organization.

Consider the Federal Government as an employer. Optimal test use would save the Government about $15.61 billion per year. However, the figure in Table 3a for the Cleary method shows that whereas majority hiring would be at 10 percent, minority hiring would be at only 1.5 percent. On the other hand, if a hiring quota were instituted, the hiring rates would be the same, but the economic loss would be 5 percent of savings or $800 million per year.

The preceding discussion was based on the assumption that cognitive ability is being used for selection. The losses are much less if psychomotor ability is the relevant predictor. Also, the difference in hiring rate is much less for psychomotor ability. With validity and selection ratio comparable to that of the Federal Government, the hiring rates for the Cleary method would be 10 percent for the majority and 6 percent for the minority. If the quota method were used, then the loss in savings would be 3 percent.

Economic disaster: the low-cutoff procedure

The most ruinous method of achieving racial balance is the method of setting a very low-cutoff score and then hiring randomly from among those who are above that score. This method is ruinous for two reasons:

(1) It eliminates nearly the entirety of the savings achieved through hiring on the basis of ability.
(2) It is inferior to the other quota methods in terms of the amount of minority hiring.

A number of different procedures exist for identifying the very low cutoff point. However for simplicity, the analysis below will consider only a typical value (though no important point is lost in this assumption). The low cutoff will be assumed to be chosen so that 80 percent of the majority applicants will "pass" the test. The minority "pass" rate will then be 52 percent.

The mathematics of the low-cutoff model are straightforward; all calculations are done as if the selection ratio were 80 percent. The cutoff is .84 standard deviations below the majority mean ability (and hence .16 standard deviations above the minority mean). The mean ability for those hired is .35 for the majority applicants and -.06 for the minority applicants.

Table 4. A comparison of the productivity losses and relative minority hiring rates for 5 methods of using ability tests for selection (selection ratio is 10 percent, validity is .50). The four methods abbreviated are those considered in Table 3: C=Cleary, T=Thorndike, D-C=Darlington-Cole, and Q=quota. The minority baseline is assumed to be 20 percent.

	Model or Scoring Method				
	C	T	D-C	Q	Low Cutoff
Percent Savings Lost	0	1	3	5	84
Relative Minority Hiring Rate*	16	44	68	100	52

*Relative minority hiring rate is defined as $\frac{\text{Percent Minority Hired}}{\text{Percent Majority Hired}}$

Table 4 presents a comparison of five different methods of personnel selection using ability tests: the four "models" of the previous section and the low-cutoff method. The situation considered consists of a validity of .50, a selection ratio of 10 percent, and a minority baseline of 20 percent. These figures show a stark contrast between the low-cutoff method and the professionally derived scoring methods. As a procedure for guaranteeing minority hiring, the method is poor; it is approximately equal to the Thorndike method and distinctly inferior to the Darlington-Cole and quota methods. Economically, the low-cutoff procedure is a complete disaster; 84 percent of the benefit of hiring on ability is lost.

It is particularly important to contrast the effects of the quota model with those of the low-cutoff method. The low-cutoff method has been sold to

employers as a way of getting around quotas. Yet the economic losses for the low-cutoff method far exceed those for the use of quotas. The quota method leads to hiring minority applicants at the same rate as for majority applicants, yet the quota method yields productivity losses of only 5 percent while the low-cutoff loss rate is 84 percent. Thus the quota method is superior to the low-cutoff method on both economic grounds and on the basis of racial balance.

Table 5 presents a comparison of four different methods of using an ability test to select workers for the Federal Government: random hiring, the low-cutoff method, the quota method, and optimal selection (i.e. ranking). The situation is assumed to be: validity of .50, selection ratio of 10 percent, and a minority applicant population of 20 percent. Table 5 shows the low-cutoff method to be only slightly better than random hiring in terms of dollars saved in production costs, or in terms of hiring workers with promotion potential. The low-cutoff method is better than random hiring in terms of weeding out workers so poor that they create special problems; it reduces the number of such workers by about half. The quota method improves over random hiring and over the low-cutoff method by a dramatic amount on all economic dimensions. Furthermore, the quota method is far superior to the low-cutoff method in terms of increasing minority hiring. The quota method does introduce noticeable loss on any economic dimension, though not nearly the loss entailed with the low-cutoff method.

Table 5. A comparison of four different methods of using an ability test to select entry level workers into the Federal Government (validity of .50, selection ratio of 10 percent, minority baseline of 20 percent).

	Method of Selection			
	Random	Low Cutoff	Quota	Optimal or Ranking
Annual Savings in Billions of Dollars	0	2.50	14.83	15.61
Percent Hired with Promotion Potential	8.8	11.7	29.2	39.4
Percent Hired of Very Poor Workers	12.4	6.6	1.2	.7
Relative Minority-Hiring Rate (percent)	100	52	100	16

The analysis above is clear. Why is the Federal Government mandating an EEO policy which is both an economic disaster and an inferior method of improving minority hiring? The low-cutoff method is a disaster by any ethical or economic standard. The key question for policy analysis is this: In what economic areas can the United States afford to reduce productivity by the amount required to use quotas to create racial balance?

POTENTIAL AND ACTUAL ECONOMIC BENEFITS FOR ORGANIZATIONS WHICH HIRE THROUGH THE U.S. EMPLOYMENT SERVICE

The dollar value to employers who use the U.S. Employment Service can be figured by treating the U.S. Employment Service as a proxy employer. In 1980, the U.S. Employment Service placed 4,022,019 applicants in jobs. Average job tenure in the United States is currently about 3.6 years, and average annual wages in the jobs served by the U.S. Employment Service is about $16,220.

Potential economic benefit to employers

Average validity for optimal test use was found by Hunter (Note 5) to be .48. Informal inquiry suggests that the U.S. Employment Service typically has jobs for only about 1 in 10 of the applicants; i.e. a selection ratio of about 10 percent. Thus, if the U.S. Employment Service used tests in an optimal way, the potential dollar savings in labor costs to the participating employers would be about

$$U = N \; T \; r_{xy} \; s_y \; \bar{x}$$

$$= (4{,}022{,}019)(3.6)(.50) \quad (.40)(16{,}220) \quad (1.76)$$

= $79.36 billion per year of hires.

The difference between this figure and that for the Federal Government as an employer stems from the fact that the government only hires about 460,000 workers per year, while the U.S. Employment Service is placing over 4,000,000 people.

Furthermore, this figure does not include the benefit corresponding to elimination of very poor workers, increasing the promotion pool, and increasing the quality of work.

Actual productivity gains for placements

Unfortunately, the U.S. Employment Service does not use tests in an optimal way. The Service departs from optimal use in three ways:

(1) Informal inquiry suggests that tests are used with fewer than 10 percent of the applicants.
(2) Selection is based on the results of small sample validation studies rather than validity generalization based on the entire data bank.
(3) Recommendations are based on the low-cutoff method rather than ranking (or some other method of assuring racial balance such as quotas).

The 90 percent of the applicants who are placed without consideration of tests are placed on the basis of "counseling" which consists primarily of acquiring data on training and experience. Empirical evidence concerning the validity of training and experience as predictors of job performance has been reviewed by Beardsley (Note 7) and by Johnson, Guffey, and Perry (Note 8). These reviews both found that the empirical evidence shows training and experience ratings to be useless, average validity is actually negative, though not significantly different from 0. Thus, the use of counseling to place applicants is equivalent to random selection as far as economic benefit

is concerned. Thus the figure of $79.36 billion per year must be immediately reduced by 90 percent. That is, because of the lack of use of tests, the maximum potential savings to employers is $7.94 billion per year.

Empirical validation results based on small-sample studies are known to be strongly effected by random sampling error. Thus they will lead to regression equations that are considerably less valid than those which use validity generalization as a data base. However, no one has yet quantified the extent of such loss. This would be particularly difficult in the case of the U.S. Employment Service since they have used multiple-cutoff procedures rather than standard multiple regression. I estimate the loss due to poor methodology to about 5 percent. If my estimate is correct, then it would reduce the potential benefits of U.S. Employment Service placements to (.95)(7.94) = $7.54 billion per year.

The disastrous consequences of using the low-cutoff method have already been described. The particular low cutoff used by the U.S. Employment Service is ambiguous. Some state offices recommend applicants only if they have an "H" rating on the relevant composite for the job in question. This is equivalent to hiring randomly from among the top two thirds of the ability distribution. However, other states recommend placement for either an "H" or an "M" rating. This is equivalent to hiring at random from the top 80 percent. Hiring randomly from the top two-thirds yields a loss of 70 percent of savings. Hiring randomly from the top 80 percent yields a loss of 84 percent of savings. If the states split about 50-50 on this issue, then the net loss of savings would be about 77 percent. This reduces the potential savings from $7.54 billion per year to about $1.73 billion per year.

Thus the actual economic benefit to employers who hire through the U.S. Employment Service is not $79.36 billion per year, but $1.73 billion per year; a slippage of some 98 percent due to nonoptimal use of the GATB. To look at it more optimistically, there is a potential gain of $77.63 billion per year in benefit to employers (and hence ultimately to consumers as well) stemming from a change to optimal procedures in using ability tests to make placements. This potential increase of $77.63 billion per year in citizen benefit can be broken into two parts: an increase from $1.73 billion to $7.94 billion due to changing the procedures for using test scores, and an increase from $7.94 billion to $77.63 billion due to increasing the use of tests from 10 percent of applicants to 100 percent of applicants.

The change in how test scores are used is practically free. There are two steps. First, officials in the national office have to admit that current practices are wrong. Second, new documentation for optimal procedures must be written, tested, and distributed to state offices. This one-shot cost might come to $100,000. But that cost of $100,000 would bring an increased benefit to employers of $6.21 billion per year.

The change in the number of tests to be administered is more expensive. The problem is that the GATB is not entirely a paper and pencil test; the finger and manual dexterity tests require almost individual supervision. A clerical person can run a group of four applicants at once, and the manual part of the test takes about 15 minutes. Thus the test administrator can run about 16 applicants an hour or about 32,000 applicants per year. Since this job can be done by the lowest level clerical person, the salary should not come to much more than $7,000.00 per year, or $0.22 per GATB. If the Service were to administer 40-million GATBs per year, then the cost would come to about $8.75 million per year. This $8.75 million per year would purchase an increase in American business productivity of $7.94 to $79.36 billion per year. That is, $8.75 million would purchase $71.80 billion in benefits.

CONCLUSION

The U.S. Employment Service now saves American business about $1.73 billion per year in reduced labor costs due to improved productivity from hiring higher ability workers. This could be raised to $7.94 billion per year by changing current procedures for using tests. By abandoning the current use of the low-cutoff scoring method, the increase in work force quality would generate all but 5 percent of this increase. The remaining 5 percent would come from using validity generalization to determine prediction equations instead of small sample studies. An even greater increase from $7.94 to $79.36 billion per year could be obtained by using tests for all placements. However, this would probably require an increase in Employment Service funding of about $8.75 million per year.

Reference Notes

1. Hunter, J.E. Test validation for 12,000 jobs: an application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB). Report to the U.S. Employment Service, 1980.

2. Helme, W.E.; Gibson, W.A.; and Brogden, H.E. An empirical test of the shrinkage problems in personnel classification research. Adjutant General's Office: Personnel Research Board Technical Report, Note 84, 1957.

3. Hunter, J.E. An analysis of validity, differential validity, test fairness, and utility for the Philadelphia Police Officers Selection Examination prepared by the Educational Testing Service. Report to the Philadelphia Federal District Court, Alvarez vs. City of Philadelphia, 1979.

4. Hunter, J.E. Cumulating results across studies: a critique of factor analysis, MANOVA, and statistical testing. Invited address, American Psychological Association, New York, 1979.

5. Mack, M.J.; Schmidt, F.L; and Hunter, J.E. Dollar implications of alternative models of selection: a case study of park rangers. Unpublished manuscript available through Frank Schmidt, Office of Personnel Management, Washington, D. C. 20415.

6. Hunter, J.E. and Schmidt, F.L. Noncompensatory aspects of the utility of personnel selection instruments. Paper presented at the Midwestern Society of Multivariate Psychology, Chicago, 1978.

7. Beardsley, V.A. A study of the rating of education and experience as an examination method in the Pennsylvania State Civil Service Commission. Commonwealth of Pennsylvania, State Civil Service Commission, Bureau of Examinations, Research and Special Projects Division, 1976.

8. Johnson, J.C.; Guffey, W.L.; and Perry, R.A. When is a T and E rating valid? Paper presented to the International Personnel Management Association Assessment Council, Boston, 1980.

References

Bartlett, C.J., Bobko, P., Mosier, S.B., and Hannon, R. Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. Personnel Psychology, 1978, 31, 233-241.

Boehm, V.R. Differential prediction: A methodological artifact? Journal of Applied Psychology, 1977, 62, 146-154.

Brogden, H.E. On the interpretation of the correlation coefficient as a measure of predictive efficiency. Journal of Educational Psychology, 1946, 37, 65-76.

Brogden, H.E. When testing pays off. Personnel Psychology, 1949, 2, 171-183.

Callender, J.C., & Osburn, H.G. Testing the constancy of validity with computer generated sampling distributions of the multiplicative model variance estimate: Results for petroleum industry validation research. Journal of Applied Psychology, 1980, 65, 543-558.

Cleary, T.A., and Hilton, T.I. Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of educational measurement, 1968, 5, 115-124.

Cole, N.S. Bias in selection. Journal of Educational Measurement, 1973, 10, 237-255.

Cronbach, L.J., & Gleser, G. Psychological tests and personnel decisions. Urbana: University of Illinois Press, 1965.

Darlington, R.B. Another look at "culture fairness." Journal of Educational Measurement, 1971, 8, 71-82.

Hawk, J. Linearity of criterion-GATB aptitude relationships. Measurement and Evaluation in Guidance, 1970, 2, 249-251.

Hunter, J.E. Construct validity and validity generalization. In Proceedings of the Construct Validity Colloquiom. Princeton, N.J.: Educational Testing Service, 1980.

Hunter, J.E., and Schmidt, F.L. A critical analysis of the statistical and ethical implications of five definitions of test fairness. Psychological Bulletin, 1976, 83, 6, 1053-1071.

Hunter, J.E., and Schmidt, F.L. Fitting people to jobs: Implications of personnel selection for national productivity. Chapter to appear in E.A. Fleishman (Ed.), Human Performance and Productivity, 1981, in press.

Hunter, J.E., Schmidt, F.L., and Hunter, R. Differential validity of employment tests by race: A comprehensive review and analysis. Psychological Bulletin, 1979, 86, 721-735.

Hunter, J.E., Schmidt, F.L., and Rauschenberger, John M. Fairness of psychological tests: Implications of four definitions for selection utility and minority hiring. Journal of Applied Psychology, 1977, 62, 245-260.

Katzell, R.A., & Dyer, F.J. Differential validity revived. Journal of Applied Psychology, 1977, 62, 137-145.

Lent, R.H., Aurbach, H.A., and Levin, L.S. Research design and validity assessment. Personnel Psychology, 1971, 24, 247-274.

O'Connor, E.J., Wexley, K.N., & Alexander, R.A. Single-group validity: Fact or fallacy? Journal of Applied Psychology, 1975, 60, 352-355.

Pearlman, K., Schmidt, F.L., and Hunter, J.E. Validity generalization results for tests used to predict training success and job proficiency in clerical occupations. Journal of Applied Psychology, 1980, 65, 373-406.

Schmidt, F.L., and Hunter, J.E. The future of criterion-related validity. Personnel Psychology, 1980, 33, 41-60.

Schmidt, F.L., and Hunter, J.E. Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 1977, 62, 529-540.

Schmidt, F.L., Berner, J.G., and Hunter, J.E. Racial differences in validity of employment tests: Reality or illusion? Journal of Applied Psychology, 1973, 53, 5-9.

Schmidt, F.L., Gast-Rosenberg, I., and Hunter, J.E. Validity generalization results for computer programmers. Journal of Applied Psychology, 1980, 65, 643-661.

Schmidt, F.L., Hunter, J.E., and Caplan, J.R. Validity generalization results for two jobs in the petroleum industry. Journal of Applied Psychology, in press.

Schmidt, F.L., Hunter, J.E., and Pearlman, K. Task differences and validity of aptitude tests in selection: A red herring. Journal of Applied Psychology, in press.

Schmidt, F.L., Hunter, J.E., and Pearlman, K. Progress in validity generalization: comments on Callender and Osburn and further developments. Journal of Applied Psychology, in press.

Schmidt, F.L., Hunter, J.E., and Urry, V.W. Statistical power in criterion-related validity studies. Journal of Applied Psychology, 1976, 61, 473-485.

Schmidt, F.L., Hunter, J.E., McKenzie, & Muldrow, T. The impact of valid selection procedures on workforce productivity. Journal of Applied Psychology, 1979, 64, 609-626. (b)

Schmidt, F.L., Hunter, J.E., Pearlman, K. & Shane, G.S. Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. Personnel Psychology, 1979, 32, 257-281. (a)

Taylor, H.C. and Russell, J.T. The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. Journal of Applied Psychology, 1939, 23, 565-578.

Thorndike, R.L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

☆ U.S. GOVERNMENT PRINTING OFFICE: 1983 - 381-608 - 814/6922